Week 17 : Distributions & Sampling Strategies

pending tasks	
type	

The focus is on continuous RV , normal distribution and sampling strategies.

Continuous Random Variable

> **Recap :** PMF talks about the distribution of the probabilities of RV.

Cumulative distribution function quantifies the probability of $P(X \leq x)$. It has values of RV on x-axis and probabilities on y-axis.

- E.x : rolling two dice

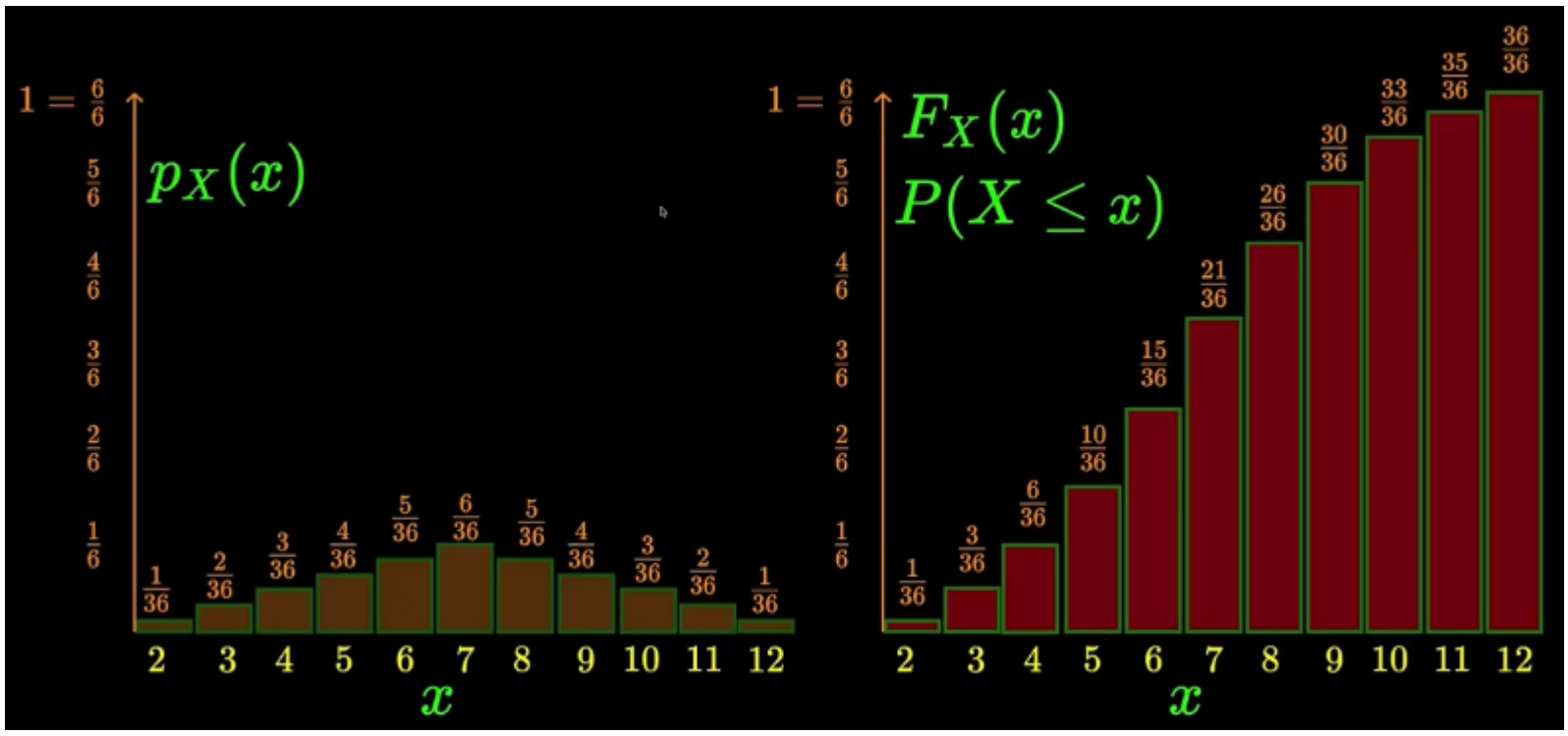


Fig.1 PMF and CDF of rolling two dies,

- Total probability (sum of PMF for a given RV) = 1 (unit mass)
- PMF shows the share of unit mass that each value of RV takes. In the above example of dies, the $x = 7$ takes $\frac{1}{6}$ of the unit mass (total probability).
- If a random variable can take **infinite values** (as in the case of continuous RV), the share of unit probability that each value takes is 0, unlike in the case of discrete RV.

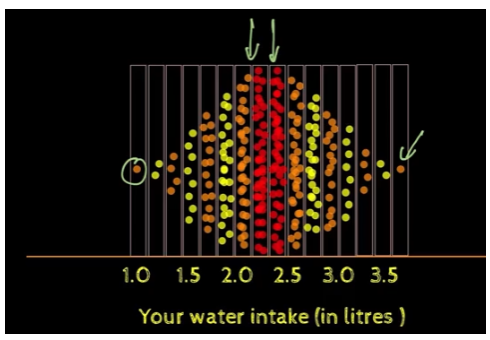


Fig.2a An example of water intake (a continuous RV) to depict distribution using intervals.

- If the intervals are made narrow, the density plot around points can be made. **This plot does not talk about the probability at a particular point but the density around that point.**
- For continuous RV, talking about the probability in ranges makes more sense than single-discrete values. Thus, the possible values are split into ranges of RV occurrence to realise the distribution of points. Fig.2 shows a person's water intake distribution 512 days.

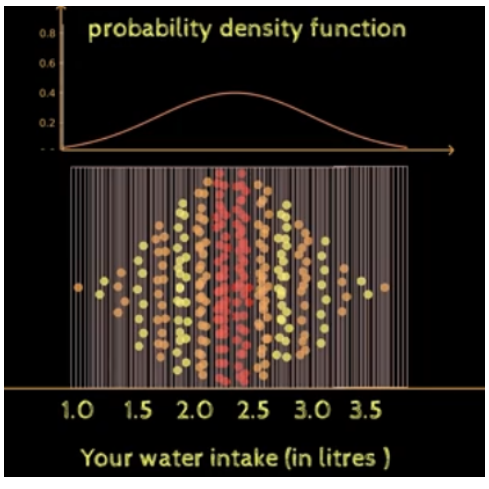


Fig.2b The plot for PDF around a RV point.

Intuition : Density vs Mass

- PDF - mass / unit length : a linear density function.
- PDF is denoted using $f_X(x)$and should satisfy the properties$\int_{-\infty}^{+\infty} f_X(x)dx = 1$ and $f_X(x) \geq 0$,where X is the RV and x is the values it can take.

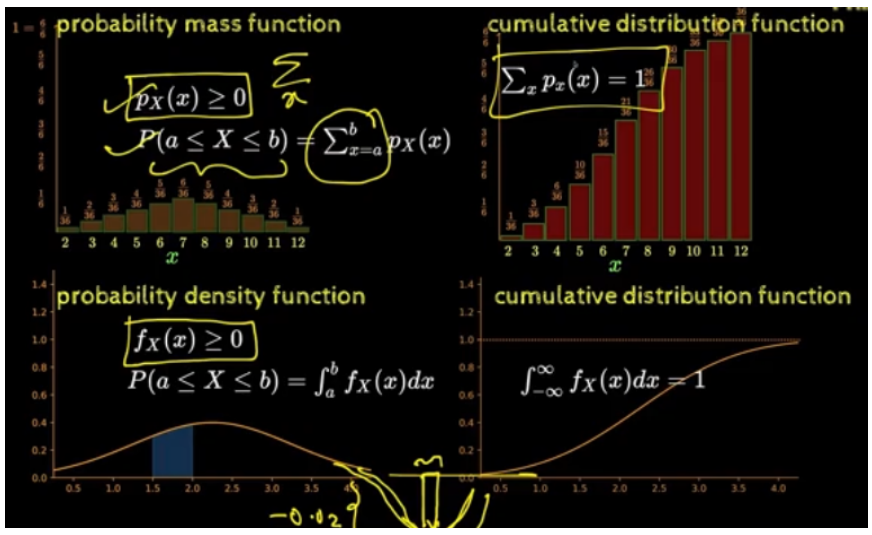


Fig.3 Summarising PMF vs PDF.

- The summation in the discrete case is integral in the continuous case.
- Density is the area under the curve that gives the mass of an interval, quantifying the probability of that interval.

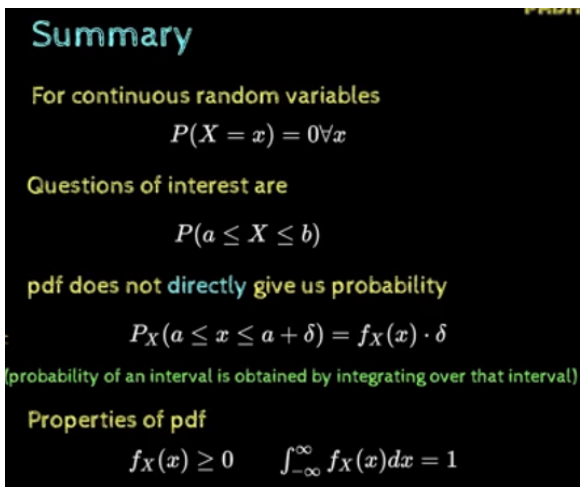


Fig.4 Summarising PDF.

Uniform Distribution (Continuous)

- In the discrete case the RV variable could take values between a and b and, each value had an equal probability $\frac{1}{b-a+1}$ where, $b - a + 1$ is the number of elements between a and b.

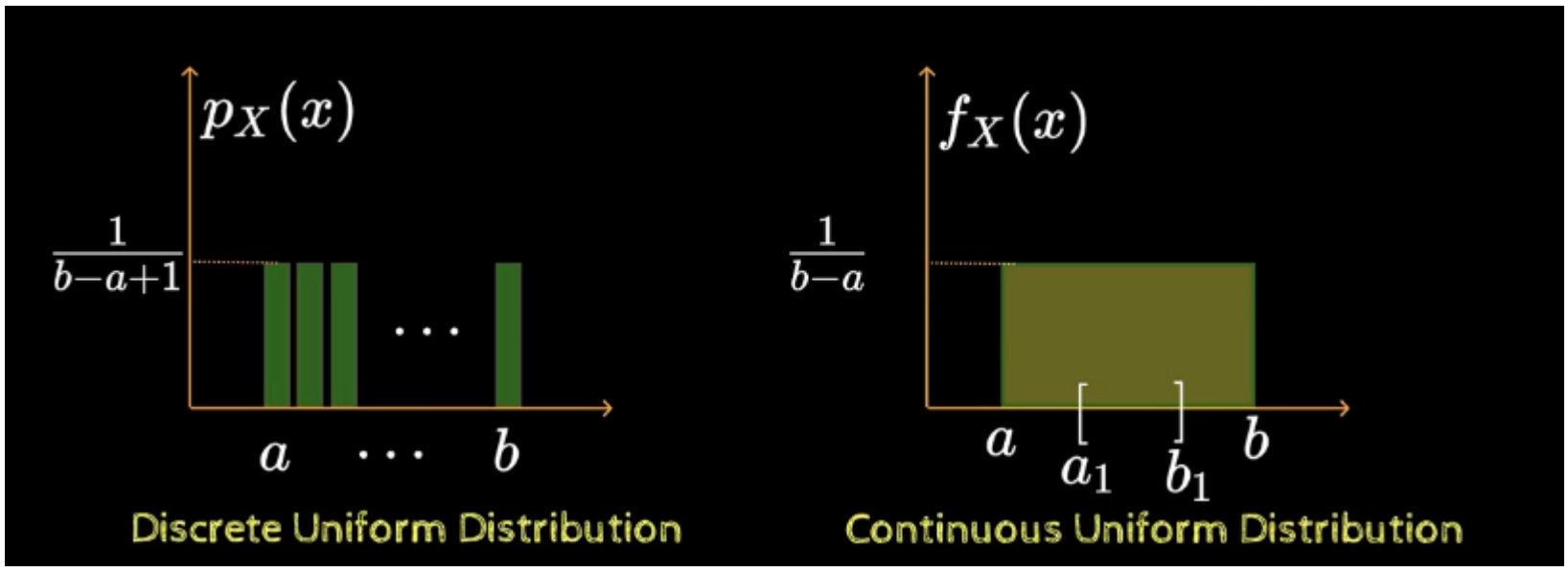


Fig.5 Uniform distribution in discrete(mass function) and continuous case(density function).

- The continuous uniform distribution is a rectangle with length $b-a$ and the height should also be equal to $\frac{1}{b-a}$ to satisfy the first property of PDF. Thus, for any range (a_1, b_1) the probability can be easily calculated and is given by

$$P(a_1 \leq X \leq b_1) = \frac{1}{b-a}(b_1 - a_1)$$

- The modes in the data can be interpreted from the RV vs PDF plot.
- The expectation of a continuous RV is given by $E[X] = \int_{-\infty}^{+\infty} x f_x(x) dx$. For uniform distribution it is equal to $\frac{a+b}{2}$.
- The expectation of a function of continuous RV is given by $E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$. For uniform distribution it is qual to $\frac{a^2+ab+b^2}{3}$.
- The variance of the continuous RV is given by $Var(X) = E[X^2] - (E[X])^2$. For uniform distribution it is equal to $\frac{(b-a)^2}{2}$.

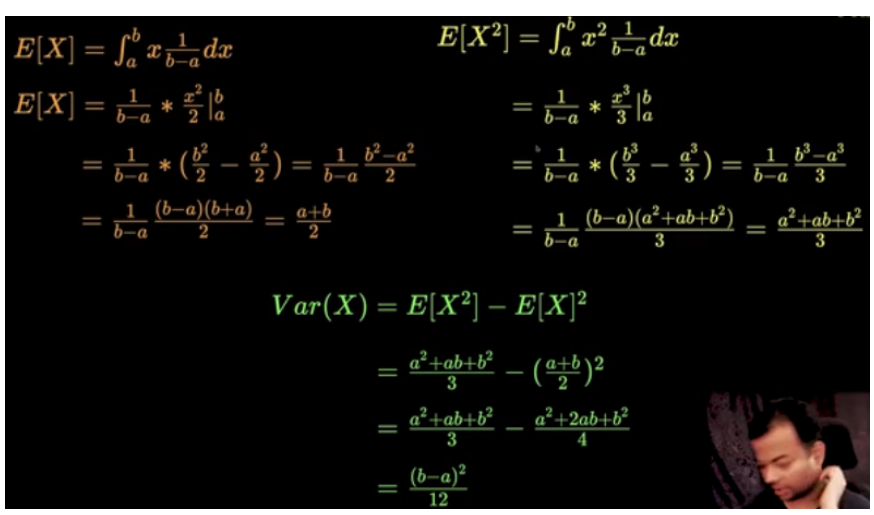


Fig.6 Calculating mean and variance for uniform distribution.

Some Fun with Functions

- An intuition about a formula can be derived by plotting the function. This helps in understanding the distribution of functions and their relative rate of growth.
- Squaring a function, introduces symmetry in plot and the rate of growth will be relatively higher.

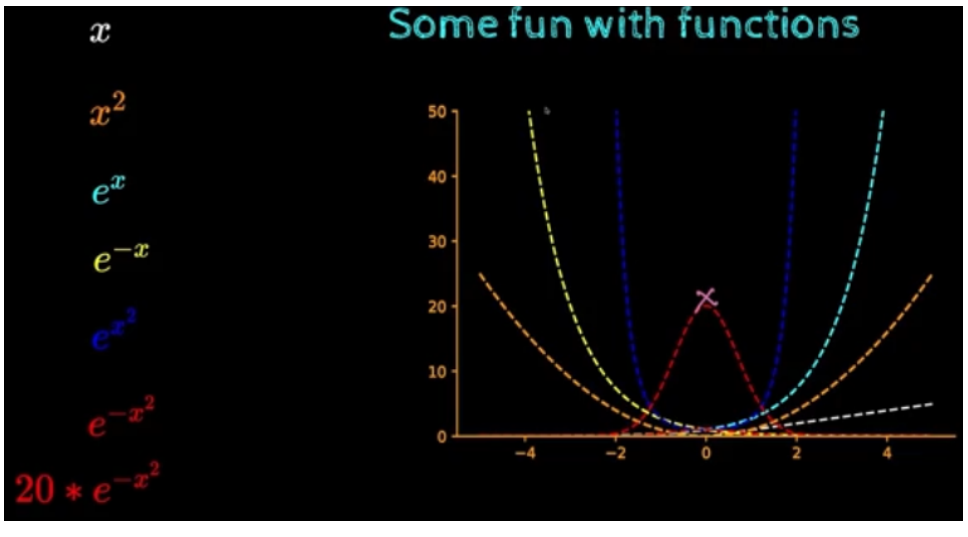


Fig.7 Examples of plotting functions.

Normal Distributions

- The function e^{-x^2} is symmetric about $x = 0$.
- The bell shape is observed in many real time scenarios. The point of peak and spread may be different but a similar curve is observed. Thus, this distribution is called a normal distribution.
- Examples can be the disribution of runs scored, cholesterol levels observed in a population, the height of students in a class, noise observed in recieved transmissions, marks scored by the students and noise in the data used for ML.
- The normal density function is given by

$$N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2}$$

as seen it is always greater than zero.

- The integral of the function is 1.

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2} dx = 1$$

Thus, proving this is a valid density function.

Probability Density Function

- The expectation $E[X]$ of normal density function is 0.
- $Var[X] = 1$
- Thus, the representation of zero mean (μ) , unit variance (σ^2):

$$N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2} dx$$

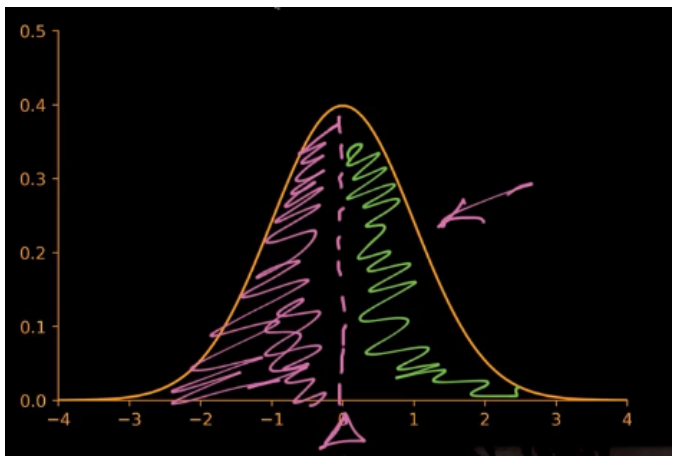


Fig.8.1 fulcrum-balance example of expectation of normal function.

- When the mean is not 0 i.e. the function peaks at a point other than 0, then the formula for density function is

$$N(\mu,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} dx$$

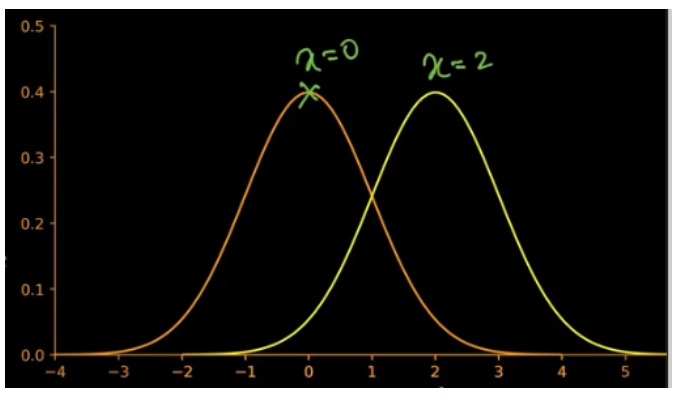


Fig.8.2 An example of shift in point when x=2.

- Change in the spread of the curve (variance) can be achieved by varying the denominator of the power of e. This essentially modifies the rate of change of the function, thus, achieving varying spread.

$$N(0,\sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}\, dx$$

smaller σ(<1) results in steeper curve and larger values (>1) result in a curve more spread out.

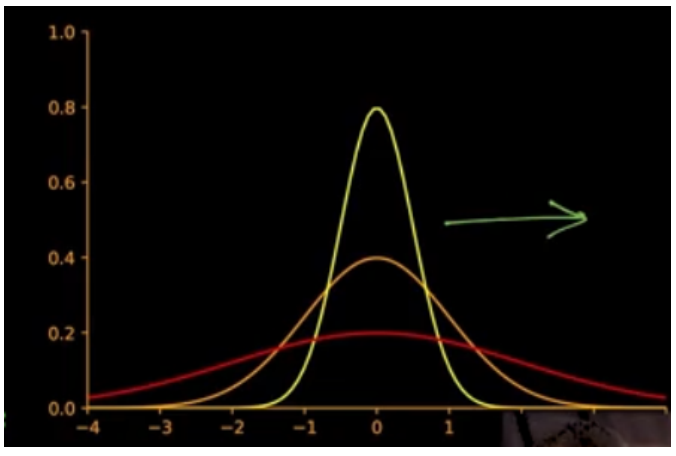


Fig.8.3 An example of varying the spread using variance.

- Thus, the general form of normal density function is :

$$N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx$$

Standard Normal Distribution

- For standard normal distribution ($\mu = 0, \sigma^2 = 1$), the area within 1σ of the mean constitutes 68% , area within 2σ constitutes 95% and area within 3σ constitutes 99.7% of the area under the curve. This is known as the **68-95-99.7 rule**.

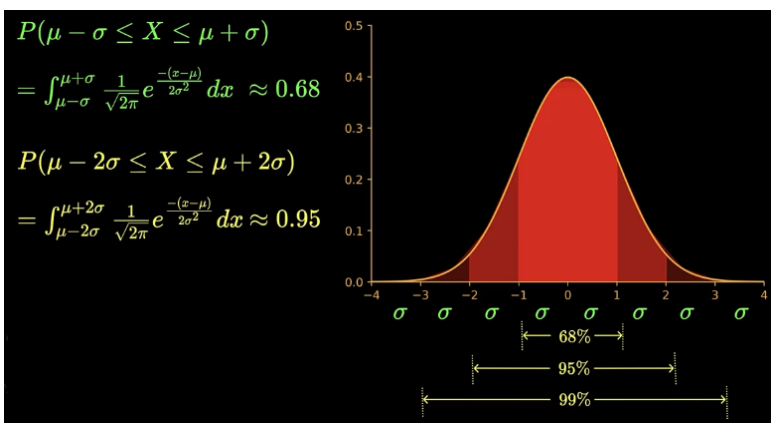


Fig.9 The 68-95-99.7 rule.

Sampling Methods

def unbiased samples - (t = 15:35)

> **Recap : Population** is the total collection of all objects that we are interested to study. **Sample** is a subgroup of the population that we study to draw inferences about the population.

- There are two ways of sampling that results in a bias : **Convenience sampling** and **voluntary sampling(non-response bias).** Biased samples are not true representations of the population.
- Examples of convenience samples discussed are a. taking the students from first row of the class to measure average height. b. sampling people visiting the hospitals to test for covid-19. c. the election results survey conducted by Chicago The Daily Tribune in 1948 when Truman won. d. Many people born in Dec were drafted higher than the other months. All these examples involve bias in the sample introduced due to convenience.
- People who strongly believe in one way generally are willing to respond than the moderates. This is because the moderates are comfortable with the current arrangements and see no point in responding to surveys. Thus, samples which involve people participating voluntarily involve bias.
- **A voluntary or non-responsive bias occurs when people who respond/participate are very different from those who do not.** An example would disgruntled citizens are more likely to post as compared to citizens who are happy

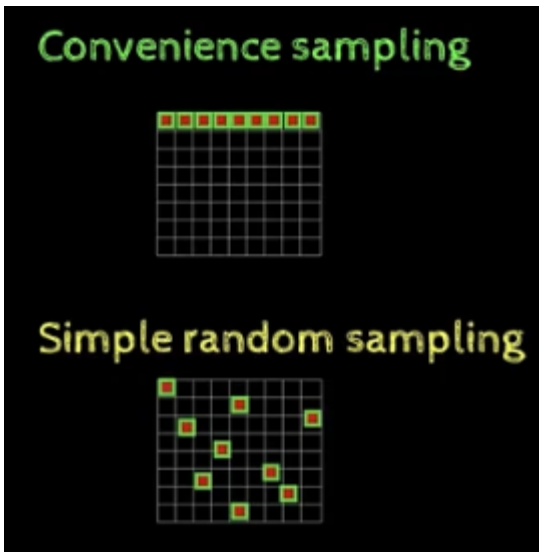
with a policy, thus, checking social media posts to predict portion of people favoring a policy would be biased.

- An **unbiased sample** is representative of the entire population and gives each element an **equal chance** of being chosen.

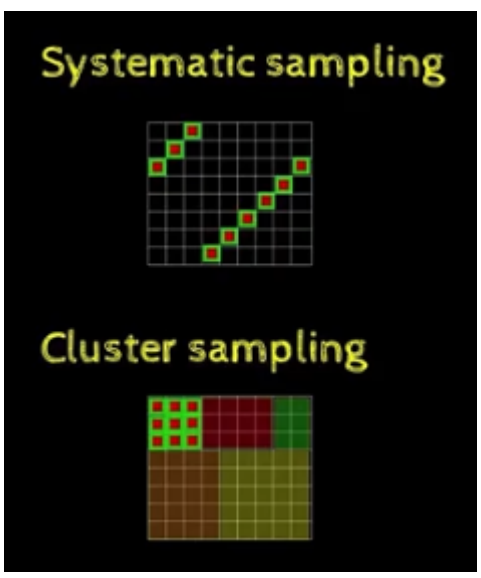
- There are five general sampling strategies:

  Convenience sampling : selecting the rows that are relatively easy to reach.

  Simple random sampling (SRS): Each element of the population has an equal chance of being selected.



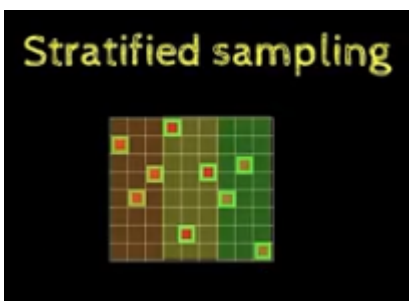

```
#creating random sample of size integers
import numpy as np
low,high,size = (1,72,9)
np.random.randint(low,high,size)
```



Stratified sampling : The population is divided into strata and samples are formed by selecting equal number of elements from each stratum by using SRS. There maybe a need to balance the uneven distribution in strata while sampling

Systematic sampling : The first element is randomly chosen and the remaining elements are chosen systematically. This works only if there is no pattern in the original dataset.

Cluster sampling : The dataset is divided into clusters based on certain attributes and each cluster constitutes a sample.



- The most common sampling strategy is SRS.

Experimental Studies

- There are two types of studies, namely observational and inferential studies. In the former the conclusions are drawn using a passive study(observation) of the data. In inferential studies, the experimenter interferes in the study to draw conclusions.
- An experiment means to apply treatment to experimental units and make observations to compare treatments.
- **Terminologies associated with an experiment**

  Response variable : the output that is to be tracked to draw conclusions.

  Experimental unit : It is the physical entity which can be assigned at random to a treatment. Any two experimental units must be capable of receiving different treatments.

  Factors : These are the attributes that can influence the response variable.

  Factor levels : The distinct values in each factor.

  Treatment : It is a combination of factors and factor levels.

Lurking variables : Attributes that aren't considered as factors but can influence the response variable. These have to be identified and controlled by either randomly assigning people to groups or all the experimental units must have similar parameters for lurking variables.

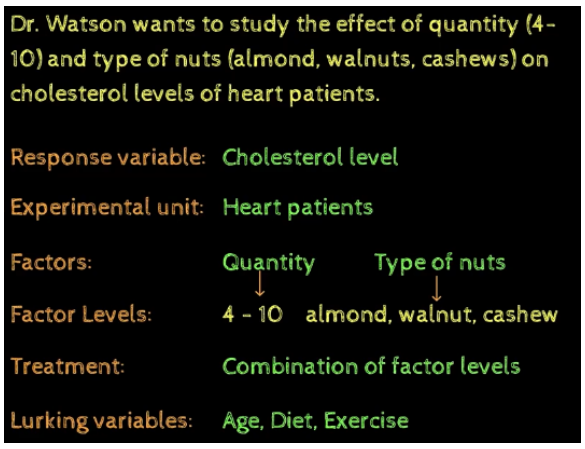


Fig.11 An exam

- A good experiment must follow the basic principles of **randomization, repetition and control.** Wherein, the subjects are randomly assigned to treatment groups, multiple subjects are given the same treatment and the effect of lurking variables have to be controlled.
- **Placebo effect** is the psychological effect a treatment has on a patient/subject. To study this the actual treatment is given to the experimental group and a placebo is given to the control group. None of the participants of the study know which group they belong to.
- There are two types of experiments: single blind experiments and double blind experiments. In single blind experiments the subjects do not know which group they belong to. In a double blind experiment, both the experimenter and the subjects do not know which group one belongs to. There is an intermediate person who knows this data.

Summary

- If a random variable can take **infinite values** (as in the case of continuous RV), the share of unit probability that each value takes is 0. A range of values is considered while defining probabilities.
- The plot for probability density function depicts the probability density around a point belonging to the random variable and not the probability at the particular point.
- PDF is denoted using $f_X(x)$and should satisfy the properties$\int_{-\infty}^{+\infty} f_X(x)dx = 1$ and $f_X(x) \geq 0$,where X is the RV and x is the values it can take.
- For uniform distribution in a given range (a, b) the expectation is equal to $\frac{a+b}{2}$, for function $E[X^2]$ is $\frac{a^2+ab+b^2}{3}$ and the variance is $\frac{(b-a)^2}{2}$.
- An intuition about a formula can be derived by plotting the function. This helps in understanding the distribution of functions and their relative rate of growth.
- Normal distribution is the bell curve observed in many real world datasets. It has the general form $N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$
- The standard normal distribution follows a 68-95-99.7 rule.
- Convenience sampling and response sampling result in biased sampling. Convenience sampling, SRS, systematics sampling, cluster sampling, stratified sampling are the general five sampling methods.
- In a good experiment allows for randomization, repetition and control.

MCQ

1. on a test, the students' scores are distributed normally. The 16th percentile is 110 while the 98th percentile is 350. What is the mean sore in this test?
    1. 190
    2. 300
    3. 250

4. Cannot determine.

2. For a standard normal distribution, the value of mean is
    1. ∞
    2. 1
    3. **0**
    4. not defined
3. The total area under the standard normal curve is
    1. **1**
    2. ∞
    3. cannot be determined
    4. depends on the equation of the curve
4. Skewness of normal distribution is
    1. Positive
    2. negative
    3.
    4. Undefined
5. Which of the following random variables is expected to be discrete?
    1. The weights of mechanically produced items.
    2. **Number of children in a classroom**
    3. The half life of elements
    4. The distance between Delhi and other state capitals
6. Let $X \sim N(3,22)$. What does this tell us about the distribution of X ?
    1. X is binomial with n = 3 and p = 2
    2. **X is normal with mean 3 and variance 4**
    3. X is normal with mean 3 and variance 2
    4. X is binomial with mean 2 and variance 9
7. Suppose X is normally distributed with mean 5 and standard deviation 0.4 . Using the standard transformation $Z = X - \mu \sigma$ we find $P(X \leq X0) = P(Z \leq 1.3)$. What is the value of X0 ?
    1. 6.9
    2. 4.48
    3. 2.0
    4. **5.52**